BMC Genomics

RESEARCH Open Access

Fine mapping of RNA isoform diversity using an innovative targeted long-read RNA sequencing protocol with novel dedicated bioinformatics pipeline

Camille Aucouturier[1,2,3], Nicolas Soirat[1,2,4], Laurent Castéra[1,2], Denis Bertrand[4], Alexandre Atkinson[1], Thibaut Lavolé[1], Nicolas Goardon[1,2], Céline Quesnelle[1], Julien Levilly[1], Sosthène Barbachou[1], Angelina Legros[1], Olivier Caron[5], Louise Crivelli[6], Philippe Denizeau[7], Pascaline Berthet[8], Agathe Ricou[1,2], Flavie Boulouard[1,2], Dominique Vaur[1,2], Sophie Krieger[1,2,3†] and Raphael Leman[1,2*†]

Abstract

Background Solving the structure of mRNA transcripts is a major challenge for both research and molecular diagnostic purposes. Current approaches based on short-read RNA sequencing and RT-PCR techniques cannot fully explore the complexity of transcript structure. The emergence of third-generation long-read sequencing addresses this problem by solving this sequence directly. However, genes with low expression levels are difficult to study with the whole transcriptome sequencing approach. To fix this technical limitation, we propose a novel method to capture transcripts of a gene panel using a targeted enrichment approach suitable for Pacific Biosciences and Oxford Nanopore Technologies platforms.

Results We designed a set of probes to capture transcripts of a panel of genes involved in hereditary breast and ovarian cancer syndrome. We present SOSTAR (iSofOrmS annoTAtoR), a versatile pipeline to assemble, quantify and annotate isoforms from long read sequencing using a new tool specially designed for this application. The significant enrichment of transcripts by our capture protocol, together with the SOSTAR annotation, allowed the identification of 1,231 unique transcripts within the gene panel from the eight patients sequenced. The structure of these transcripts was annotated with a resolution of one base relative to a reference transcript. All major alternative splicing events of the *BRCA1* and *BRCA2* genes described in the literature were found. Complex splicing events such as pseudoexons were correctly annotated. SOSTAR enabled the identification of abnormal transcripts in the positive controls. In addition, a case of unexplained inheritance in a family with a history of breast and ovarian cancer was solved by identifying an SVA retrotransposon in intron 13 of the *BRCA1* gene.

†Sophie Krieger and Raphael Leman contributed equally to this work.

*Correspondence:
Raphael Leman
r.leman@baclesse.unicancer.fr

Full list of author information is available at the end of the article

Conclusions We have validated a new protocol for the enrichment of transcripts of interest using probes adapted to the ONT and PacBio platforms. This protocol allows a complete description of the alternative structures of transcripts, the estimation of their expression and the identification of aberrant transcripts in a single experiment. This proof-of-concept opens new possibilities for RNA structure exploration in both research and molecular diagnostics.

Keywords RNA splicing, Long read sequencing, HBOC, Isoform assembly, Automatic annotation

Background

Splicing of pre-mRNAs is a major source of transcript diversity. Also known as alternative splicing, this mechanism concerns at least 90% of multi-cassette genes [1]. Three main mechanisms lead to this transcript diversity: the exon skipping, the use of alternative splice sites and intron retention (Figure S1). As an example for the human gene *KCNMA1*, more than 500 transcripts were described [2, 3]. The knowledge of this diversity is a crucial step to understand and explore physiological and pathological processes [4–6]. Moreover, a variety of genomic variations could affect RNA splicing. Indeed, 3.8% of genomic variations from the Exome Aggregation Consortium (ExAC) had an impact on splicing [7]. The study of RNA transcripts also resolves the functional impact of genomic variation in the context of inheritance disease [8–10]. In addition, alternative splicing could interfere with the clinical interpretation of genomic variations [11–13].

Over the past decade, several new methods based on short read sequencing (SRS) partially dealt with these challenges. High-throughput RNA sequencing [14], in particular targeted short read RNA sequencing [15, 16] and high-throughput minigene splicing assay [7, 17] have been developed and widely used. Despite the advantage of these technologies, they are limited in exploring the complete structure of isoforms [18]. Then, the recent demonstration of the value of long-read sequencing (LRS) in describing the structure of isoforms represents a significant advance in the field [19]. The two main technologies to perform this sequencing are provided by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio). Both platforms are based on single molecule sequencing but differ in nucleotide detection. Briefly, ONT sequencing platform uses the variation in ionic current within the pores of an impermeable membrane, while PacBio platform synthetizes the complementary DNA strand with a fluorescent-labeled nucleotide.

RNA long read sequencing by ONT or PacBio platforms was performed either on direct RNA/mRNA [20] or total cDNA. However, this approach does not provide a comprehensive collection of isoforms, especially for genes with low expression levels. The second known approach used long range RT-PCR amplicons [21, 22] but limits targeting to one or a few genes with the potential addition of a major PCR bias. Recent methods using hybridization RNA capture combined with long read sequencing were developed. However, they were used for different applications: long non-coding RNA [23–25], single cell RNA [26] or single nucleus RNA [27].

Therefore, we described a new protocol of target enrichment by probes for coding mRNA long read sequencing compatible with both ONT and PacBio sequencing platforms, including a dedicated pipeline of transcript analysis. In this study, we designed probes to capture transcripts from a panel of genes involved in hereditary breast and ovarian cancer (HBOC) syndrome. In addition, we developed and evaluated a dedicated bioinformatics pipeline named SOSTAR for "iSofOrmS annoTAtoR", to assemble isoforms from ONT sequencing. Using the novel annotation tool, SOSTAR provides the ability to annotate isoforms, regardless of the assembly method used and generates a "human" comprehensive annotation of the isoform structure. We validated our protocol by sequencing negative controls from healthy donors, and positive controls from patients carrying *BRCA1* spliceogenic variants.

Methods

Sample collection

Lymphoblastoid cell lines from 8 patients having signed consent were established for this work. Four samples were negative controls, i.e. healthy donors, without any molecular abnormalities. These healthy donor samples were collected as part of the CASOHAR (*CAncer du Sein et/ou de l'Ovaire Héréditaire – ARN*, N°ID-RCB 2015-A00598-41) clinical protocol. Two patients were positive controls carrying complex variants identified by targeted DNA sequencing on an Illumina NextSeq 500. One control carries a duplication of *BRCA1* exon 8, and the other an intronic genomic deletion in *BRCA1* intron 15. Several criteria (Table S1) were defined to select highly predisposed families for RNA sequencing to investigate unexplained hereditary causes. An initial screening by targeted short read RNA sequencing was performed on eleven families (Table S2) according to these criteria. Two probands from one of these families with abnormal splice junctions of the *BRCA1* gene were added to our cohort to investigate the unexplained inheritance in this family.

Probes design

A custom panel of 28 genes (Table S3) was designed using xGen Lockdown probes from Integrated DNA

Technologies (IDT). Transcript structures and sequences described in the RefSeq database [28] were used for the design. Initially, the probe positions were automatically defined with a tailing of 0.1x to 1x in steps of 0.1 according to the transcriptomic sequence of the gene panel (Supplementary material). To reduce the risk of overselection of a particular isoform, overlapping probes between two contiguous exons were removed. We arbitrarily set a minimum number of 5 probes per gene. Thus, in the second step, probes for genes below this threshold were added to reach this number. These new probes were from previous designs with a higher tailing rate. The percentage of GC and the specificity of the probe sequence to the target were then checked according to IDT recommendations. Finally, a set of 367 probes was used to capture our panel, including, for example, ten probes for the *BRCA1* gene (Figure S2).

Library preparation

Specifically, for PacBio platform, the AMPure beads (Beckman, #A63881) required a washing according to the following protocol. From 500 µL of AMPure beads, the supernatant was saved after bead precipitation on a magnetic rack. The beads were washed 5 times with 1 mL of nuclease free water plus one more wash with 1 mL of Qiagen elution buffer (Qiagen, #19086). They were then resuspended with the saved supernatant.

RNAs were extracted from cells using the RNeasy Plus Mini Kit (Qiagen, #74134) according to the manufacturer's instructions. All samples had a RNA Integrity Number (RIN) above 9.

First strand cDNA synthesis was carried out with the SMARTer PCR cDNA Synthesis Kit (Clontech #634925) according to the manufacturer's protocol, from 1 µg of total RNA. The resulting first strand product was diluted to 150 µL of H_2O and used for large scale PCR.

For each sample, sixteen PCR reactions were performed using the PrimeSTAR GXL DNA Polymerase (Clontech #R050A). Thermal cycling conditions were 98 °C for 30 s, followed by 12 cycles of 98 °C for 10 s, 65 °C for 15 s, and 68 °C for 10 min, and a final extension of 68 °C for 5 min. Amplicons were separated into 2 fractions for purification. Fraction 1 was purified twice using 1X washed AMPure beads and fraction 2 once using 0.4X washed AMPure beads. The two fractions were quantified using the 2200 TapeStation system (Agilent) and then pooled to obtain an equimolar pool of 1 µg of DNA.

The libraries were hybridized according to the IDT manufacturer's recommendations using the specific blockers (5' AAG CAG TGG TAT CAA CGC AGA GTA C 3') and (5' TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT / 3InvdT/3') at 1 mM.

Two post-hybridization PCR reactions were performed using Takara LA Taq DNA Polymerase Hot-Start version (Clontech, #RR042A) for each sample. Thermal cycling conditions were 95 °C for 2 min, followed by 14 cycles of 95 °C for 20 s, and 68 °C for 10 min, and a final extension of 72 °C for 10 min. Amplicons were then purified using 1X washed AMPure beads.

After DNA repair, barcodes were ligated using the SMRTbell Library Construction and Sequencing kits (PacBio) according to the manufacturer's protocol (PN 101-070-200 Version 05 [November 2017]). SMRTbell libraries were multiplexed and loaded onto Sequel II to get the subreads sequencing.

End repair and dA tailing were performed using the NEBNext Ultra™ End Repair/dA-Tailing module (New England BioLabs #E7546S). Barcoding using the PCR Barcoding Expansion kit (#EXP-PBC096) was performed with the SQK-LSK109 ligation sequencing kit (ONT) according to the Nanopore community protocol. Samples were washed twice with Short Fragment Buffer (ONT) before a final elution in 15 µL of elution buffer (ONT). Samples were quantified using Qubit® Fluorometer (ThermoFisher Scientific).

ONT sequencing was performed on the MinIon Mk1B device and PacBio sequencing on a Sequel II device. Prepared libraries were multiplexed and loaded on MinION flow cell (R9.4.1) according to instructions. MinKNOW software (v21.06.0) was used for running the MinION sequencer during 64 h. Additional flush buffer was added as the number of pores used decreased. (Fig. 1A).

Targeted short read RNA sequencing

Targeted short read RNA sequencing was performed on lymphoblastoid cell lines from patients from eleven families selected according to the protocol described by Davy et al. [16]. The data was analyzed using the SpliceLauncher tool [29] to detect aberrant splice junctions. Loss of gene expression was investigated using the DESeq2 tool [30].

The four negative control samples and the two probands were sequenced from the same RNA extraction on the same gene panel using this approach. Splice junctions detected by LRS were verified with those obtained by SRS. Only splice junctions supported by at least 10 reads were retained for further analysis.

Bioinformatics analysis

PacBio Iso-Seq

Highly accurate long reads (HiFi reads) were generated from PacBio sequencing using the Iso-Seq v3 pipeline (https://github.com/PacificBiosciences/IsoSeq) with default parameters. HiFi reads were aligned to version 46 of the human reference genome assembly (GRCh38). Only high quality isoforms were considered for analyses.

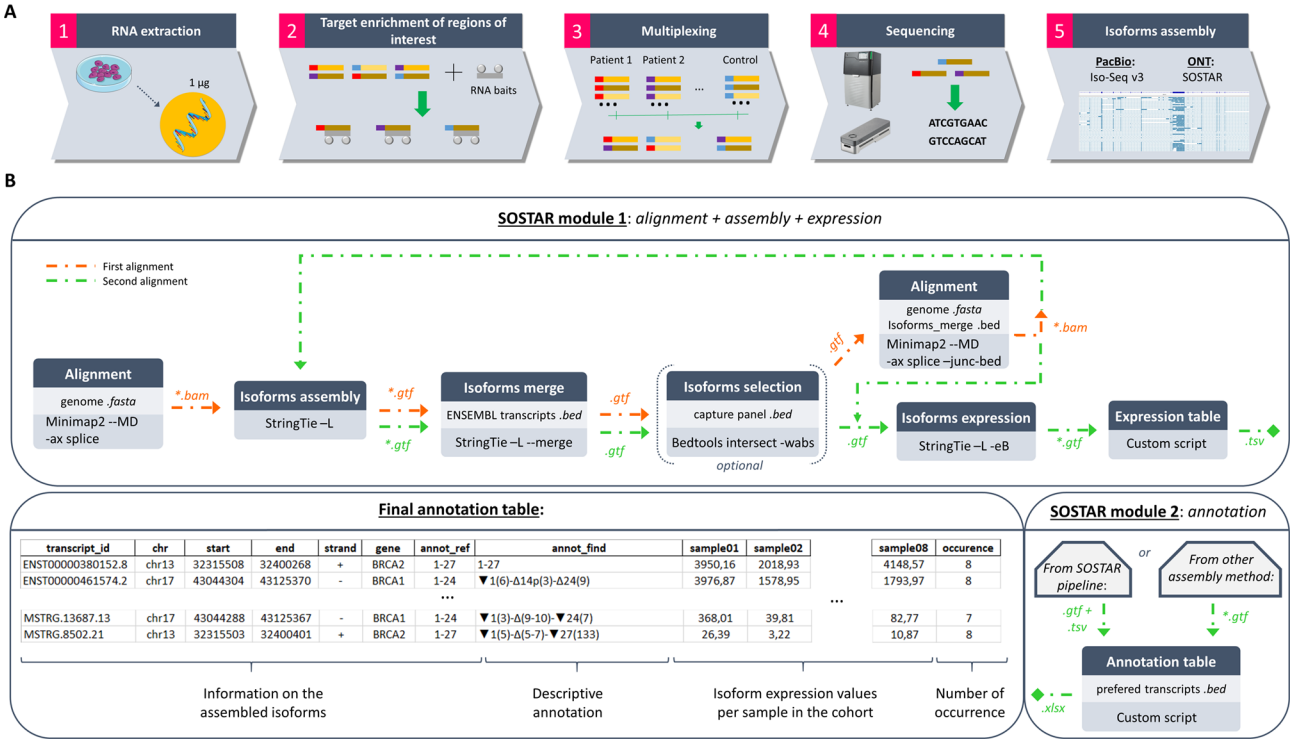

Fig. 1 Targeted long read RNA sequencing workflow. **(A)** Overview of the sequencing protocol from cell lines to isoform assembly **(B)** Description of the SOSTAR pipeline

ONT SOSTAR

ONT basecalling and data demultiplexing were performed using Guppy tool (v5.0.11) with default options. These data were then processed by the SOSTAR pipeline. SOSTAR is publicly available on GitHub (https://github.com/LBGC-CFB/SOSTAR), and has been split into two modules that can be run separately (Fig. 1B).

The first module aligned ONT data, assembled isoforms and computed the isoform expression. This module uses a combination of existing and recognized tools [31] and was optimized using a range of options to produce isoforms with a good level of confidence. Alignment to version 46 of the human reference genome assembly (GRCh38) was first performed using Minimap2 [32] (v2.24) in the splice mode with the options "-MD --ax splice". Isoforms were assembled using the StringTie tool [33] (v2.0). From the bam files, StringTie was first used in the assembly mode with the minimum fraction abundance option set to 0 to assemble all possible isoforms and generate a GTF file *per* sample. These files were then merged using the StringTie "merge mode" to build a reference of all potential isoforms detected in the patient cohort. During this merge step, isoforms were annotated using GENCODE annotation. Using bedtools intersect [34] (v2.30.0), this merge file was then filtered on the 28-gene panel with "-S" option to force strand specificity. A second round of alignment was then performed using Minimap2 with the same options as before, adding the "--junc-bed" option with the isoforms from the merge file produced earlier. Using the new alignment files, isoforms were assembled as described above with the minimum fraction abundance option set to 0.01 (default) up to the new merge file. Finally, isoform expression metrics *per* patient were calculated using the StringTie expression mode. Expression metrics obtained were coverage, Fragments Per Kilobase Million (FPKM) and Transcripts Per Million (TPM).

The second module is an innovative module that uses a specially designed and written tool to describe the isoforms assembled in the first module. The tool provides a descriptive and comprehensive annotation of each isoform structure. This module is compatible with any long read technology and any assembly method. It only requires the different gtf files per patient containing the isoforms to be annotated and one gtf file containing the reference transcript for each gene (one transcript per gene). Isoforms were described relative to reference transcripts (Table S3) by an annotation including only the alternative splicing events. The alternative splicing events were annotated according to Lopez et al. [13]. This nomenclature uses "Δ" to refer to a skipping of a reference exon, "▼" an inclusion of a reference intron, "p" a shift of an acceptor site and "q" a shift of a donor site. For partial skipping or inclusion, number of skipped or retained nucleotides is indicated between brackets (Figure S1). Relative positions are indicated between square brackets.

"Exo" refers to an exonization of an intronic sequence, and "int" to an intronization of an exonic sequence with the relative intron or exon number in front. A final table of all annotated isoforms was generated. This table allowed different levels of filtering, including isoform ID, isoform coordinates, gene name, reference annotation, descriptive annotation, expression values of isoform *per* patient and number of occurrences in the patient cohort. These different levels of filtering allow the user to easily detect abnormal isoforms and quickly understand their structure.

Data Analysis

The aligned data were visualized with Integrative Genomics Viewer (IGV) software [35]. Depth coverage of the 8 patients was calculated using featureCounts tool (v2.0.6) for the 28 captured genes with "-L" option to count the long reads. This tool was used with a reference human genome assembly version 46 (GRCh38) file downloaded from GENCODE (https://www.gencodegenes.org/human/release_46.html) and filtered on the 28 captured genes. Percentages of reads on and off target were calculated using featureCount results. On target reads were calculated as the number of aligned reads assigned to the 28 genes out of the total number of aligned reads. Using the expression module of StringTie tool, expression values *per* isoform were calculated on the isoforms assembled by SOSTAR pipeline. In the patient cohort, the splicing junction expressions were calculated using formula (1) for LRS or formula (2) for SRS:

$$Expression_{i^{th}junction} = \sum_{j=0}^{n_{sample}} \left(\sum_{k=0}^{n_{isoforms}} read\_count_{ijk} \right) \quad (1)$$

$$Expression_{i^{th}junction} = \sum_{j=0}^{n_{sample}} (read\ count_{ij}) \quad (2)$$

Results

Raw sequencing data

Our library preparation yielded 40.16 ng/µL of library, sufficient for both PacBio and ONT sequencing. Average length of library fragments were 2 kb with a range length from few hundreds nucleotides (nt) to over 50,000 nt. From PacBio platform a total of 9 Gb of data were generated, representing 5 M Circular Consensus Sequencing (CCS) reads. ONT platform generated 27 Gb of 10 M reads.

Read count for the 28 genes was correlated between the two sequencing platforms with a correlation coefficient of 0.6977 (Fig. 2A). For both technologies, the *STK11* and *XRCC3* genes exhibited the lowest level of coverage. *PTEN* and *NBN* genes presented the highest coverage. On the gene panel, average reads count per gene was 12,196 reads for PacBio and 29,904 reads for ONT per sample.

On target rate estimated from read count was similar between samples (Fig. 2B) for both platforms. PacBio sequencing reached an average on target of 64%, and ONT sequencing achieved 60% of on target.

IsoSeq (PacBio) and SOSTAR (ONT) pipelines assembled a total of 89,379 and 1,231 unique isoforms in selected genes, respectively. Isoforms length distribution was investigated between these two pipelines (Fig. 2C). IsoSeq-assembled isoforms averaged 2 kb in length, while SOSTAR supported the longest isoforms. Indeed, isoforms detected by IsoSeq had a maximum length of 8,000 nt while SOSTAR reached a maximum of 12,000 nt. The limit of 12,000 nt represented the longest full length transcript of our panel for *ATM* gene (12,915 nt).

Comparison of splice junctions

Splice junctions from both SRS and LRS (IsoSeq and SOSTAR) were compared for the 28 genes of the panel. A total of 3,214 junctions were found from SRS, compared to 8,110 (PacBio) and 1,370 (SOSTAR) junctions from LRS (Fig. 3A).

Among junctions observed by SRS, 1,360 junctions were not detected by LRS but were supported by a lower average read count per sample (29.90 [1.83; 2,957.5]) compared with the 1,854 junctions identified by at least one long-read pipeline (8,740.55 [1.83; 264,893.5]). Among these common SRS-LRS junctions, SOSTAR mostly detected junctions supported by a high average read count (Fig. 3B). While PacBio mainly detected junctions with a low average read count.

Common junctions between SRS and the LRS pipelines represented 897 splicing events. Among these common junctions, physiological junctions were the most represented. All MANE select transcripts from Refseq were found for the 28 genes.

Comparison of the expression values of common junctions between SRS and LRS

Correlation between expression values of the common splice junctions from LRS (SOSTAR isoforms) and SRS was investigated. Expression values were correlated for the 28 genes panel (r=0.722, p-value<0.001) between LRS and SRS (Fig. 4).

Alternative splicing of *BRCA1* and *BRCA2* genes

Alternative splicing events of the *BRCA1* (NM_007294.4) and *BRCA2 (*NM_000059.3*)* genes were explored on the aligned long read from ONT sequencing to validate the data obtained by our protocol. Indeed, alternative splice sites of the *BRCA1* (NM_007294, OMIM#113705)

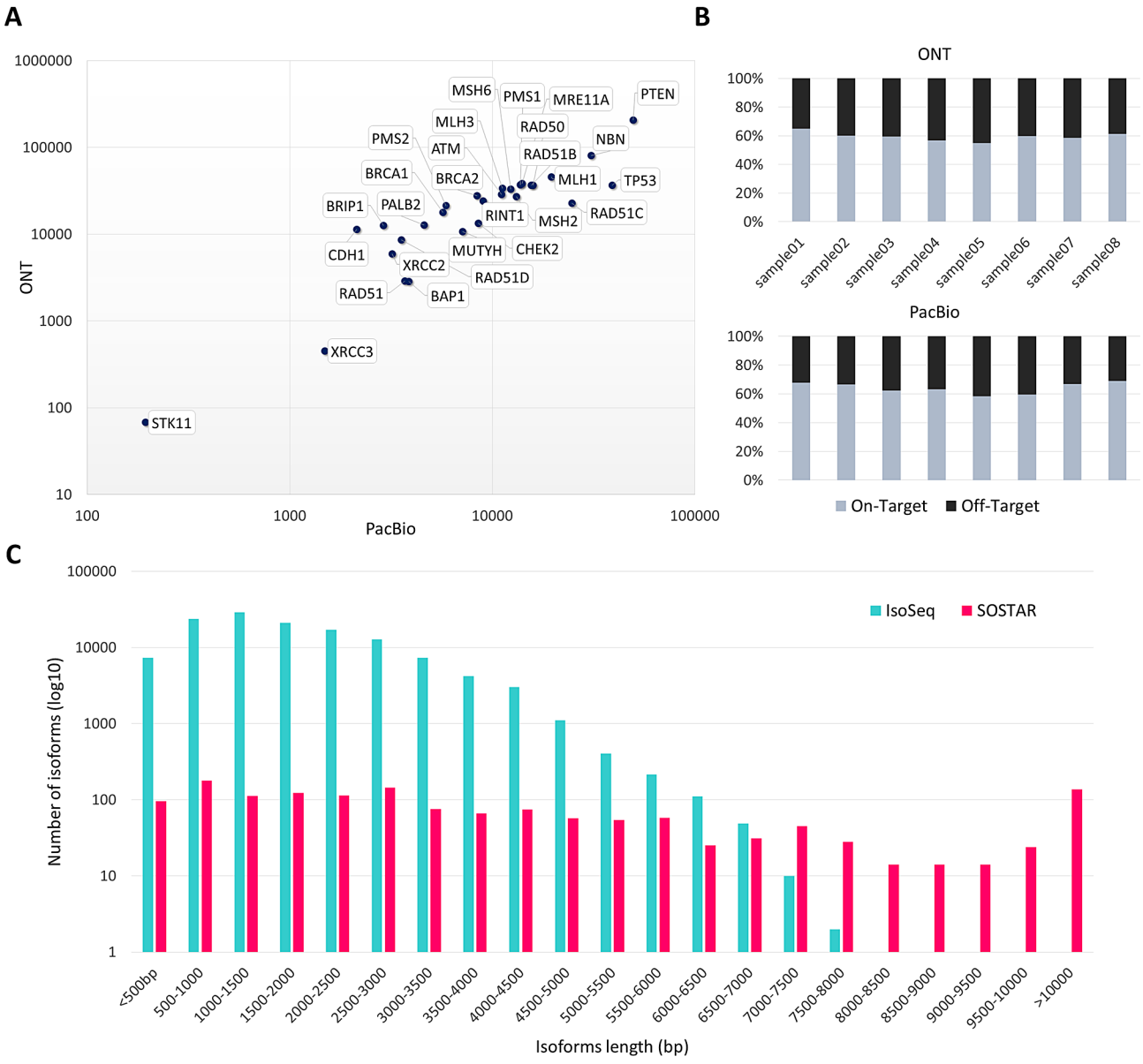

Fig. 2 Overview of the results generated by targeted long read sequencing. **(A)** Average coverage calculated for the 28 genes in the patient cohort. Values are plotted on a logarithmic scale **(B)** Percentage of on and off target rates for the 28 genes in the patient cohort **(C)** Distribution of isoform lengths assembled by the long read pipelines

and *BRCA2* (NM_000059, OMIM#600185) transcripts involved in this syndrome were well known [36, 37].

Previously, Colombo et al. [36] described 10 predominant *BRCA1* splicing events; the in-frame events: Δ1q(6), Δ5, Δ8p(3), Δ(9–10), Δ(9–11), Δ11q(3309), Δ13p(3), Δ14p(3) and the out-of-frame events: Δ5q(22), Δ9. All of these *BRCA1* events were found in our data. The major alternative splicing event found was the Δ5. The two out-of-frame splicing events were the least represented of all splicing events. Our results showed that these events co-occur (Fig. 5A). Interestingly, all pairwise combinations of these splicing events are possible, except when the 2 events are physically incompatible i.e. Δ(9–11) and Δ11q(3309). Pairwise combination of Δ5 and Δ(9–10) represented the majority of all possible pairwise splicing events. These combinations of splicing events were in-frame.

For *BRCA2* gene, four predominant alternative splicing events were described by Fackenthal et al. [37]: the Δ3, Δ(6–7), Δ12 in-frame events and the Δ(17–18) out-of-frame event. As observed in *BRCA1*, all pairwise combinations were found in the long reads (Fig. 5B). The Δ(6–7) splicing event was mostly found, and combination of Δ3 and Δ(6–7) represented the major pairwise combinations. This combination was out-of-frame.

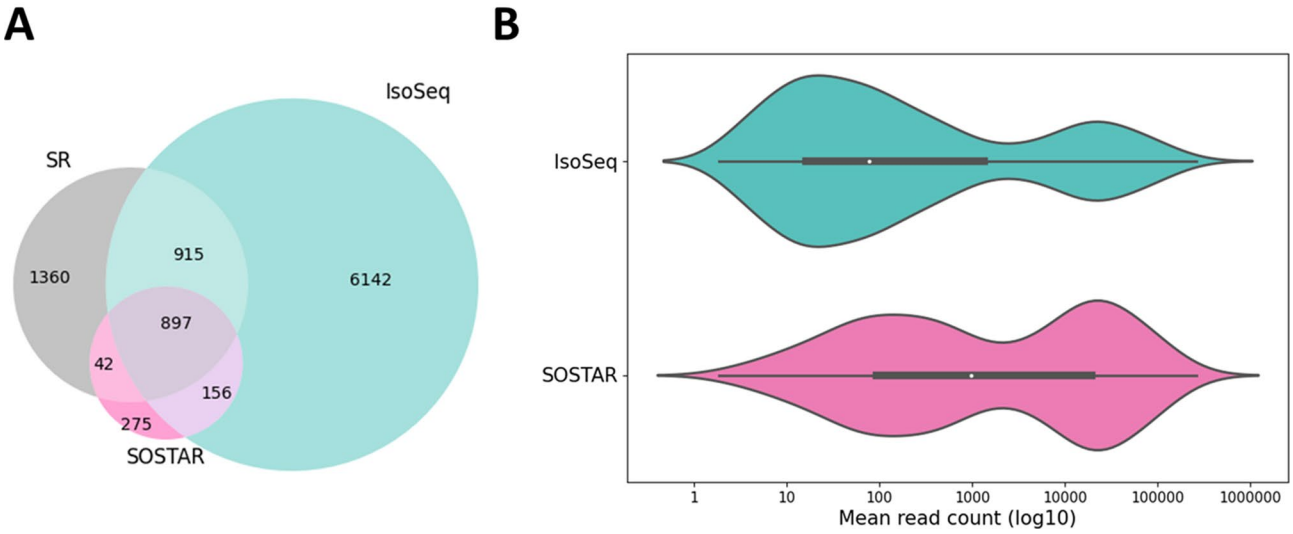

Fig. 3 Comparison of splicing junctions between SRS and LRS for the whole gene panel. **(A)** Venn diagram of splicing junctions detected by SRS and LRS **(B)** Violin plot of SR mean read counts of common junctions between SRS and LRS junctions. Values are plotted on a logarithmic scale

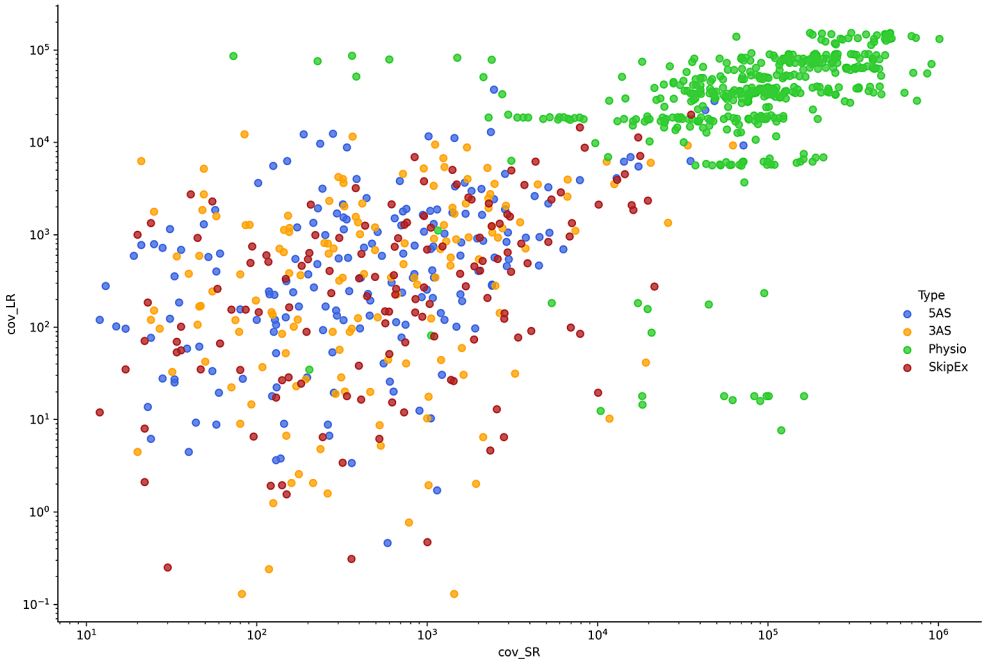

Fig. 4 Expression values of common splicing junctions between short and long read sequencing (SOSTAR isoforms). Values are plotted on a logarithmic scale. Junction are represented by types: 5AS = alternative 5′ splice site / 3AS = alternative 3′ splice site / Physio = physiological junction / SkipEx = exon skipping

Among the alternative junctions detected by SRS, the five most frequent combinations within *BRCA1* gene [▼1q(534)/Δ(9_10); Δ11q(3309)/Δ14p(3); Δ1q(6)/Δ11q(3309); Δ(9_10)/Δ11q(3309); Δ8p(3)/Δ11q(3309)] represented 25.35% (72/284) of total junction combinations. In addition, other major event were found: ▼1q(89) (Fig. 5C). This event was mainly observed with the alternative junctions previously described by Colombo et al. [36]. Several pseudo-exons were also identified in intron 3 (▼3p(4047) and ▼3q(5261) corresponding to an exon of 116 nt) and in intron 13 (▼13p(2785) and ▼13q(3070) corresponding to an exon of 66 nt).

For *BRCA2* gene, ▼20p(1327), ▼20q(4306); ▼25p(907), ▼25q(1183) and ▼24p(11650), ▼24q(2984), junctions were over represented (Fig. 5D). These junctions corresponded to pseudo-exons creation

A.

Δ1q(6)										3802
Δ5									31022	196
Δ5q(22)								1086	0	286
Δ8p(3)							4729	341	492	1234
Δ9						685	155	50	274	126
Δ(9-10)					22785	0	1602	308	18382	1203
Δ(9-11)				11922	0	0	174	58	11167	210
Δ11q(3309)			4264	0	2030	143	1329	253	577	1093
Δ13p(3)		4954	780	176	3747	69	500	110	3508	433
Δ14p(3)	9017	1723	1446	753	6381	124	975	202	6373	793
	Δ14p(3)	Δ13p(3)	Δ11q(3309)	Δ(9-11)	Δ(9-10)	Δ9	Δ8p(3)	Δ5q(22)	Δ5	Δ1q(6)

B.

Δ3				44604
Δ(6-7)			45136	43456
Δ12		43835	35998	35547
Δ(17-18)	33982	32800	26211	25816
	Δ(17-18)	Δ12	Δ(6-7)	Δ3

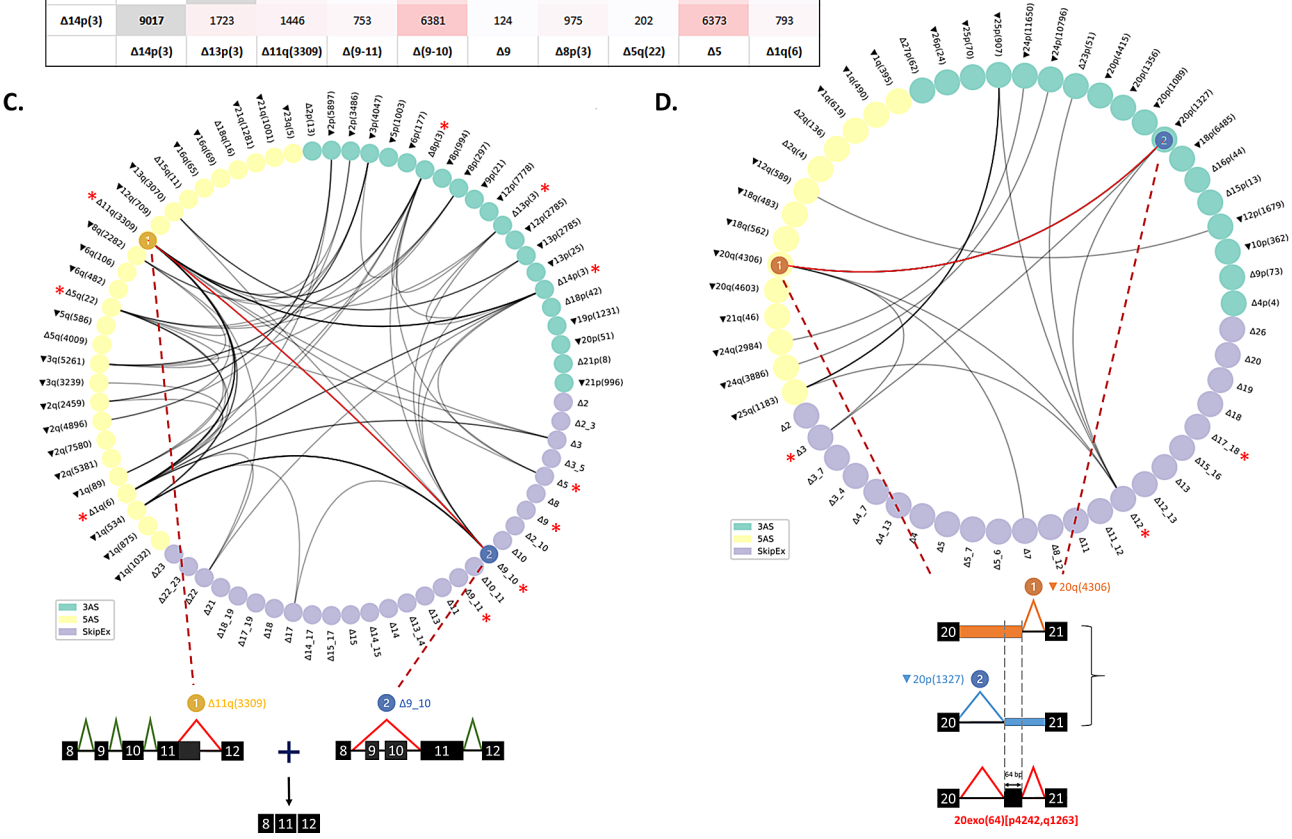

Fig. 5 Investigation of alternative splicing events in *BRCA1* and *BRCA2* genes. **(A)** Matrix of pairwise combinations of *BRCA1* splicing events within the long reads **(B)** Matrix of pairwise combinations of *BRCA2* splicing events counts within the long reads **(C)** Circos plot of common junctions between SR and LR junctions for *BRCA1* gene **(D)** Circos plot of common junctions between SR and LR junctions for *BRCA2* gene. Values are plotted on a logarithmic scale. The * represents the splicing events previously described by (Colombo et al., 2014) for *BRCA1* gene and (Fackenthal et al., 2016) for *BRCA2* gene. The thickness of the line reflects the number of long reads supporting the two junctions

in intron 20 (64 nt), intron 25 (126 nt) and in intron 24 (91 nt), respectively.

Detection of aberrant isoforms

Both technologies enabled the detection of the two positive controls. They also allowed the characterization of the effects at the transcript level. The first control carried a 28 bases genomic deletion of intron 15 of *BRCA1* (c.4676-31_4676-4del). *In silico* predictions of SPiP [38] were in favor of an abolition of the acceptor splice site. The second control was a duplication of exon 8 of *BRCA1* gene.

The BAM alignment files from LRS of these two positive controls were loaded into the IGV software to visualise the events. An intronic retention, due to the use of a new acceptor splice site 64 nt upstream the natural splice site in intron 15 was observed. The intronic retention embedded the genomic deletion (Fig. 6A). This event was out-of-frame (p.Gly1560Tyrfs*5).

For the second control, different insert sizes, corresponding to the exon 8 sequence, were observed at the end of exon 8 (Fig. 6E). Several long reads were observed supporting both the exon 8 duplication and other proximal exons (exon 7 or exon 9). The duplicated exon was spliced normally. Then, we were able to assert that this

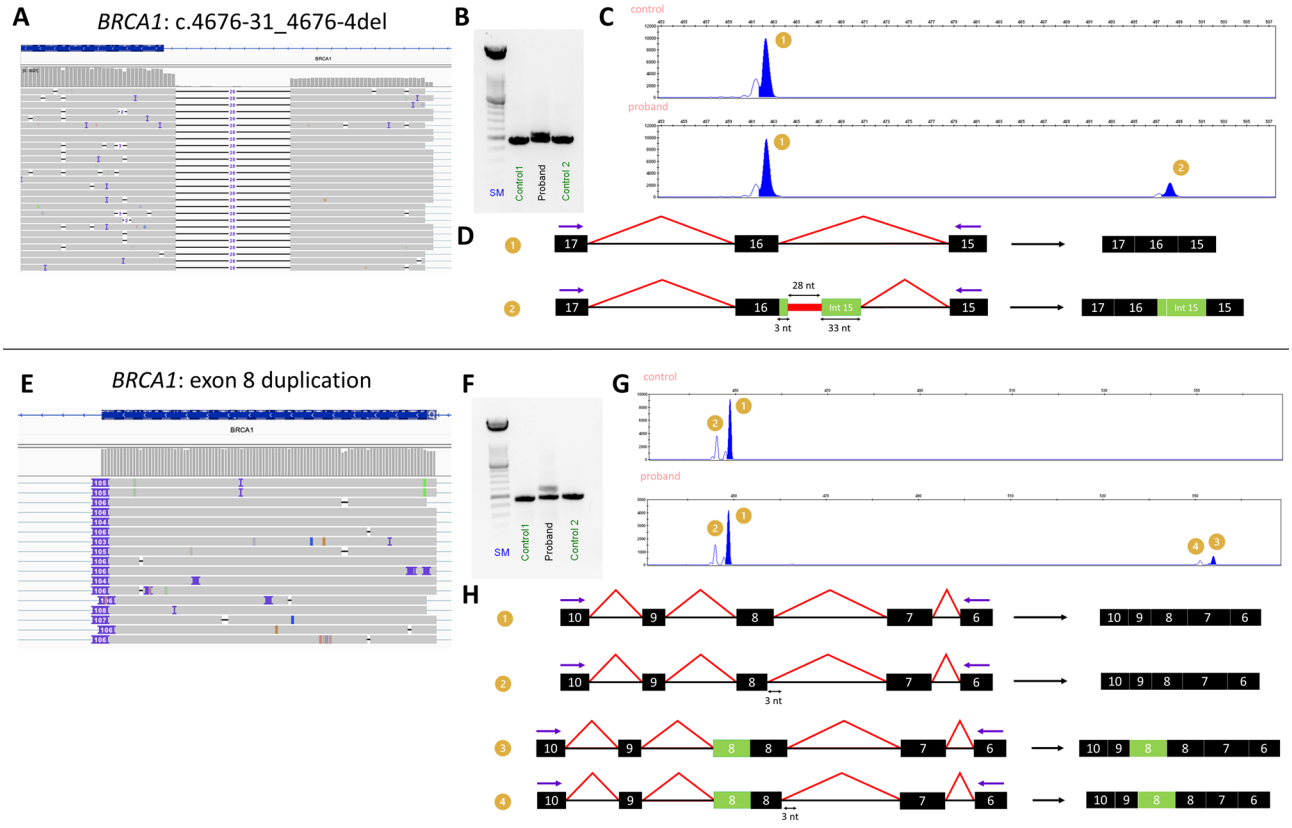

Fig. 6 Validation on positive controls. **(A)** Bam file from the LRS of the proband carrying the intronic retention in intron 15 of *BRCA1* **(B)** RT-PCR, gel electrophoresis SM: molecular weight size marker **(C)** Capillary electrophoresis of RT-PCR products **(D)** Isoform structures of fragments obtained in C. black boxes: exons; green boxes: novel exons, black thin lines: introns; red thick line: deletion; red lines: splicing junctions; purple arrows: RT-PCR primers **(E)** Bam file from LRS of the proband carrying the exon 8 duplication of *BRCA1* **(F)** RT-PCR, gel electrophoresis **(G)** Capillary electrophoresis of the RT-PCR products **(H)** Isoform structures of the fragments obtained in G

duplication was a tandem duplication, leading to an out-of-frame event.

Sanger sequencing confirmed these aberrant isoforms on specific RT-PCR amplicons (Fig. 6B-D, Figure S3, Fig. 6F-H, Figure S4) using primer pair 1 for the first control and primer pair 2 for the second control (Table S4).

Final table generated by SOSTAR pipeline allowed the complete identification and annotation of these two controls. In the first control, two isoforms carrying the aberrant event were found. According to our annotation algorithm, the aberrant event was annotated: 15exo(33) [p3028,q31]. Due to the genomic deletion, StringTie detected the use of a new exon corresponding to the 33 nt retention of intron 15 after the genomic deletion. Consequently, SOSTAR annotated this as a pseudo exon creation in intron 15. This aberrant event was assembled with previously known alternative splicing event: Δ(9–10) in one isoform and combined with the ▼1q(534) in the other. In the second control, the aberrant event was annotated: ▼8q(95). Due to misalignment, the insert size did not exactly match the 106 nt of exon 8. Two different isoforms supporting this aberrant event were assembled. Of these two isoforms, one isoform only supported this aberrant event, while the other one also supported previously known alternative splicing events: Δ(9–10) & Δ11q(3309). Indeed, the comparison of isoform expression metrics in the patient cohort allowed the identification of these aberrant isoforms for all these controls (Table S5 and Table S6).

Characterization of an unexplained heredity

The final two patients included in this analysis were two probands from a family with a history of cancer (Table S2, family number 11). The first proband (III.1) developed a breast cancer at the age of 39 and an ovarian cancer at the age of 56. Her sister (III.2) developed an ovarian cancer at the age of 49. The mother (II.2) died at the age of 51 due to breast cancer. The grandmother (I) died of a gynecological cancer (Fig. 7A). The initial targeted short read DNA sequencing [39] screening of the two probands was inconclusive.

Targeted short read RNA sequencing revealed two abnormal splice junctions using SpliceLauncher tool. These two junctions were present in both probands and absent in the other samples in the run. They were both located in intron 13 of the *BRCA1* gene

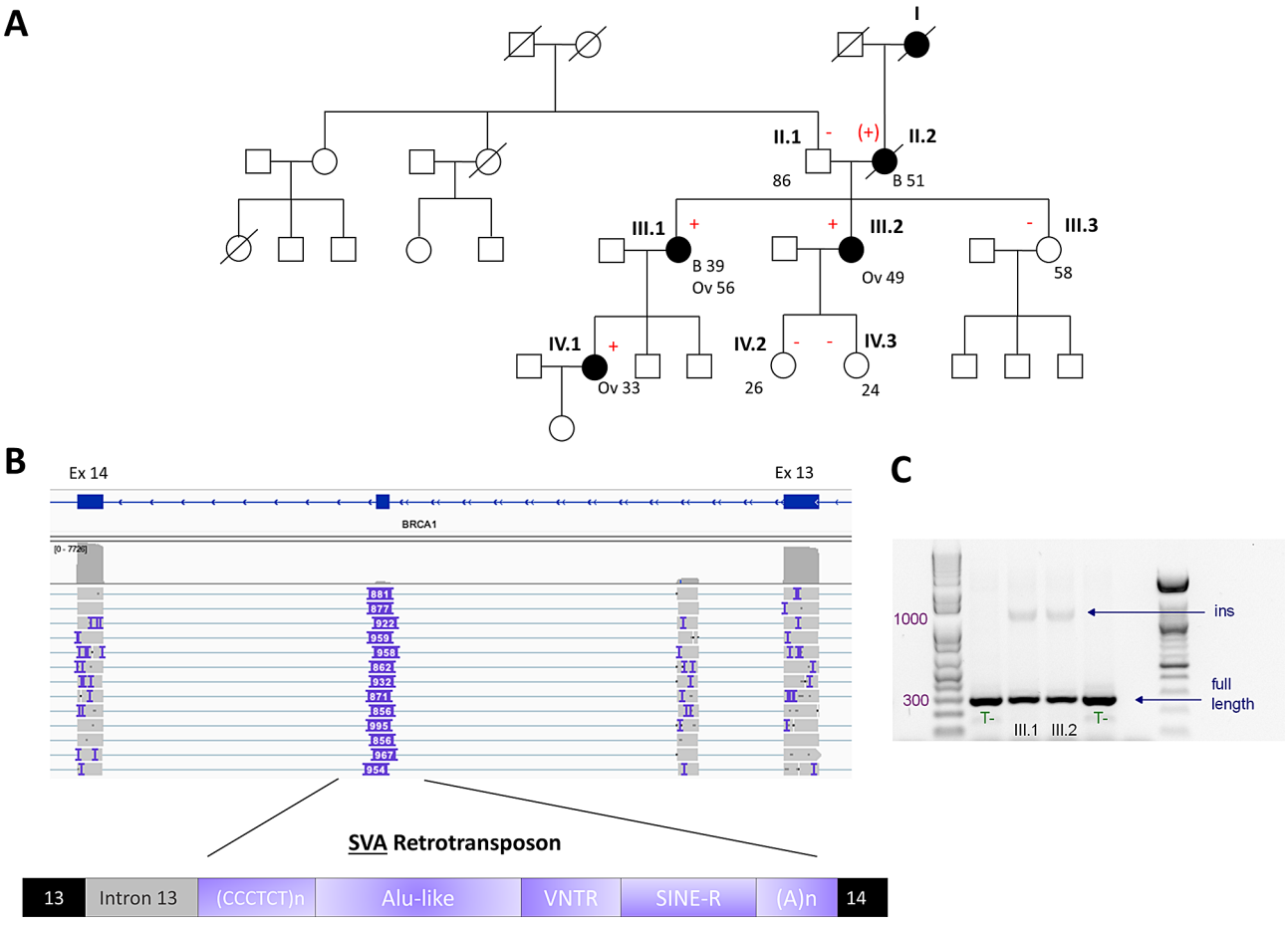

Fig. 7 An unexplained hereditary case. **(A)** Family pedigree of a family with breast and ovarian cancers. Black symbols indicate patients with breast **(B)** or ovarian (Ov) cancers. Ages are given as age at diagnosis for cancer patients and current age for living probands. The red '+' indicates the SVA retrotransposon insertion, the red '-' indicates the normal sequence at intron 13 **(B)** Bam alignment file, displayed on IGV software, from ONT long read sequencing showing the pseudo exon and the SVA retrotransposon insertion in intron 13 of *BRAC1* gene with the detailed schematic structure of the aberrant isoform **(C)** RT-PCR gel electrophoresis showing an insertion of approximately 1000 bp in cancer probands (III.1; III.2) compared to two controls (T-)

and were annotated as ▼13p(5373) starting at position c.4357+417 and ▼13q(5400) ending at position c.4358−390. These junctions were in favor of the presence of a cryptic event in this deep intronic region. The loss of *BRCA1* gene expression was calculated using the DESeq2 tool on these sequencing data.

SOSTAR pipeline assembled two different isoforms carrying an aberrant event in the *BRCA1* gene. These two isoforms were present on all the samples but with a significantly higher expression value in the probands. The aberrant event was annotated as 13exo(104) [p416,q5269] and corresponded to a pseudo-exon of 104 bp in intron 13 of *BRCA1*. Looking at the alignment files on IGV, we identified this pseudo-exon as well as an insertion of about 900 bp of a repeated sequence in intron 13 of *BRCA1* (Fig. 7B). Asking the Dfam transposable element database (https://www.dfam.org/home), we identified that this sequence corresponded to an SVA retrotransposon.

Long read sequencing was also performed on high molecular weight DNA using the adaptive sampling method. Regions of interest, comprising 153 genes, were enriched with an average coverage of 50 reads per base, compared to 5 reads per base for the off-target region. The SVA retrotransposon was found in intron 13 of the *BRCA1* gene at a length of approximately 2,700 bp. Comparison between SVA sequence from RNA and SVA sequence from DNA was performed. The RNA sequence matched a portion of the complete DNA sequence.

RT-PCR with primers located on *BRCA1* exons 13 and 15 was performed (Table S4, primer pair 3). Gel electrophoresis and Sanger sequencing of theses amplicons confirmed the insertion of a sequence of approximately 1000 bp, present in the probands and absent in the controls (Fig. 7C, Figure S5). This sequence included 102 bp of a cryptic exon, followed by the beginning of the SVA retrotransposon with the hexamer repeats. Another RT-PCR with primers located on *BRCA1* exon 11 and 14

amplified only the full-length transcript with a polymorphism at position c.3548T>C (Table S4, primer pair 4). Sanger sequencing of this polymorphism revealed the absence of the mutated allele (data not shown). Thus, this mutated allele had a total effect on splicing, producing only the abnormal transcripts.

Analysis of the ovarian tumor of the second proband (III.2) highlighted a genomic instability using the GIScar test [40]. Specific RT-PCR on this ovarian tumor with primer located specifically on the SVA retrotransposon and on exon 14 of *BRCA1* (Table S4, primer pair 5) identified the abnormal transcript. Furthermore, the sequencing of the ovarian tumour revealed a loss of heterozygosity on two polymorphisms.

A specific PCR test was designed (Table S4, primer pair 6) for the identification of this SVA retrotransposon from DNA blood samples. In a first time, the unaffected father (II.1) and sister (III.3) were tested and found to be negative. The daughters (IV.1, IV.2, IV.3) were then tested. The test was positive for one of them, she underwent a prophylactic salpingo-oophorectomy which led to the diagnosis of ovarian cancer. These results showed that this SVA insertion segregated with cancers in this family. According to the ACMG classification [41], this variant met the following arguments: PS3 (from our RNA functional studies), PM2_supporting (variant absent in the general population), PP1_moderate (variant in three patients with breast or ovarian cancer and absent in disease-free relatives). The sum of these arguments allowed us to classify this SVA insertion as likely pathogenic.

Discussion

We developed an innovative targeted capture enrichment for a panel of 28 genes involved in predisposition to breast and ovarian cancer suitable for long read RNA-seq analysis. Looking at the high percentage of on target rate, our approach allowed an enrichment of these genes. Therefore, we achieved sufficient coverage rate to detect a collection of isoform structures. The study of *BRCA1* and *BRCA2* transcripts demonstrated the reliability of these data since all well-known alternative splicing events were found with an expression level coherent with the literature data [36, 37]. Common splice junctions obtained by both LRS and SRS were the most expressed compared to junctions identified only by SRS. Additionally, we identified novel associations of these splicing events leading to novel isoforms. LRS simplified the detection of complex events such as pseudo-exon, while SRS only detected the donor and acceptor splice site separately.

Regarding the positive controls, both PacBio and ONT platforms allowed the complete identification of aberrant transcripts in a single experiment. While previous experiments, based on RT-PCR or SRS, could only detected these events indirectly and required several steps of analysis. These results highlight the advantage in time and ease of LRS. Another advantage of the target enrichment was the possibility to sequence multiple patients at once, reducing sequencing cost. As a result, decreasing costs facilitate the implementation of LRS in both research and diagnostic laboratories. Multiplexing also allowed the highlighting of aberrant isoforms by comparing the isoform expression metrics between samples.

In this study, RNA from LCL samples were used. Indeed, a sufficient RNA quality (RIN>9) is required for our RNA LRS. This constraint could be a limit for ex vivo samples such as PAXgene sample, where RNA is partially decayed. Also, the design of probes played a major role in the final isoform length. Previous designs with overlapping probes or a 1x tailing were tried but resulted in increased isoform fragmentation (data not shown). This phenomenon could be explained by mechanic constraints applied to the cDNA by two probes.

Our protocol was compatible with both long read technologies (PacBio or ONT). Nanopore sequencing allowed to sequence fragments longer than PacBio sequencing. PacBio provides an all-in-one pipeline (IsoSeq) to analyze sequencing data with read consensus generation and high-quality isoforms auto-assembly. While IsoSeq assembled more isoforms, isoforms assembled by SOSTAR pipeline were the most relevant. Indeed, isoforms identified by SOSTAR were the longest and the most expressed. This pipeline is consistent with a diagnostic approach to detect aberrant isoforms with good confidence and without background noise. The versatility of SOSTAR pipeline allows the annotation module to be used separately with any GTF file comprising isoforms assembled by other tools from any LRS data. SOSTAR facilitates interpretation of LRS data by providing a descriptive and a "human" understandable annotation of isoform structures.

SOSTAR pipeline helped us to identify a mobile element in a family with an unexplained heredity, where current techniques failed. This mobile element corresponded to an SVA retrotransposon. Characterization of this element allowed us to develop a specific PCR test suitable for genetic counselling of the family. This long read approach contributes to optimize preventive and therapeutic cares. Combination of SOSTAR and other ONT pipelines could be an interesting strategy to detect aberrant isoforms with good confidence, and a comprehensive collection of isoforms. Such an insertion of an SVA retrotransposon into intron 13 of *BRCA1* was previously described in the literature [42]. This could indicate the presence of a fragile region within the *BRCA1* gene, which could potentially be susceptible to the insertion of mobile elements.

Conclusion

In conclusion, we validated a new protocol of targeted enrichment by probes for mRNA long read sequencing suitable for ONT and PacBio platforms on both negative and positive controls. We provide a pipeline, suitable for diagnostic purpose, to detect and annotate aberrant isoforms from LRS data. This pipeline elucidates an unexplained hereditary by characterizing a complex event. This proof-of-concept offers new opportunities in RNA structure exploration for research and molecular diagnostic purposes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10741-0.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Acknowledgements

We are grateful to the Site de Recherche Intégrée sur le Cancer (SiRIC, PhD Sylvain Baulande) of Curie Institut and the Génotypage et Séquençage en Auvergne (GENTYANE, PhD Charles Poncet) of Centre de Clermont Auvergne Rhône Alpes for performing the PacBio sequencing. We also thank Nolan Talhi and Peter Verhasselt, from Integrated Dna Technologies (IDT®), for their help during the design of capture probes. Finally, we thank Bénédicte Clarisse and Alexandra Leconte from the clinical trials unit of the François Baclesse Cancer Center (Caen, Normandy), Doctor Catherine Baudouin from the Clinique du Parc (Caen, Normandy), and Doctor Marie-Yolande Louis from from the surgery unit of the François Baclesse Cancer Center for managing and enrolling patients in the CASOHAR protocol.

Author contributions

C.A. Conceptualization, data curation, methodology, validation, visualization, writing–original draft, writing–review and editing. N.S. Conceptualization, methodology. L.C. Conceptualization, methodology, validation, writing–review and editing. D.B Conceptualization, methodology. A.A Resources. T.L. Resources. N.G Resources. C.Q. Resources. J.L. Resources. S.B. Resources. A.L. Resources. O.C. Resources. L.C. Resources. P.D. Resources. P.B. Resources. A.R. Validation, resources. F.B. Validation, resources. D.V. Validation, resources. S.K. Conceptualization, methodology, validation, writing–review and editing. R.L. Conceptualization, methodology, validation, writing–review and editing.

Funding

This work was supported by a grant from the French *Cancéropôle Nord-Ouest* (CNO) n° 2018/06.

Data availability

The datasets analysed during the current study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB75906 (https://www.ebi.ac.uk/ena/browser/view/PRJEB75906). The code of the pipeline is publicly available on the SOSTAR GitHub repository (https://github.com/LBGC-CFB/SOSTAR).

Declarations

Ethics approval and consent to participate

All negative controls were from a CASOHAR (CAncer du Sein et/ou de l'Ovaire Héréditaire – ARN, N°ID-RCB 2015-A00598-41) clinical trial and were submitted to and approved by the ethics committee (NTC02560818), *Comité de Protection des Personnes CPP Nord Ouest III*. All participants gave informed consent for genetic analysis and were approved by the French Biomedicine Agency.

Consent for publication

Not applicable.

Competing interests

All authors except N.S. and D.B. declare that they have no competing interests. N.S. is employed by SeqOne Genomics for the time period October 2020 to present in the context of a public-private PhD project (CIFRE fellowship #2020/0103) partnership between INSERM and SeqOne Genomics. D.B. is employed by SeqOne Genomics as Head Bioinformatics.

Author details

[1]Laboratoire de biologie et de génétique du cancer, Département de Biopathologie, Centre François Baclesse, Caen 14000, France
[2]Cancer and Brain Genomics, FHU G4 Genomics, Inserm U1245, Normandie University, Rouen 76183, France
[3]Normandie Univ, UNICAEN, Caen 14000, France
[4]SeqOne Genomics, Montpellier 34000, France
[5]Département Médecine Oncologique, Institut Gustave Roussy, Villejuif, France
[6]Service d'Oncogénétique, Centre Eugène Marquis, Rennes, France
[7]Service de génétique clinique, Centre Hospitalier Universitaire Rennes, Rennes, France
[8]Service d'Oncogénétique, Département de Biopathologie, Centre François Baclesse, Caen 14000, France

Received: 30 April 2024 / Accepted: 28 August 2024
Published online: 30 September 2024

References

1. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing', *Nat. Genet.*, vol. 40, no. 12, pp. 1413–1415, Dec. 2008, https://doi.org/10.1038/ng.259
2. Navaratnam DS, Bell TJ, Tu TD, Cohen EL, Oberholtzer JC. 'Differential Distribution of Ca2+-Activated K+Channel Splice Variants among Hair Cells along the Tonotopic Axis of the Chick Cochlea', *Neuron*, vol. 19, no. 5, pp. 1077–1085, Nov. 1997, https://doi.org/10.1016/S0896-6273(00)80398-0
3. Rosenblatt KP, Sun Z-P, Heller S, Hudspeth AJ. 'Distribution of Ca2+-Activated K+Channel Isoforms along the Tonotopic Gradient of the Chicken's Cochlea', *Neuron*, vol. 19, no. 5, pp. 1061–1075, Nov. 1997, https://doi.org/10.1016/S0896-6273(00)80397-9
4. Bonnal SC, López-Oreja I, Valcárcel J. Roles and mechanisms of alternative splicing in cancer — implications for care. Nat Rev Clin Oncol. 2020;17. https://doi.org/10.1038/s41571-020-0350-x. 8, Art. 8, Aug.
5. Park E, Pan Z, Zhang Z, Lin L, Xing Y. The Expanding Landscape of Alternative Splicing Variation in Human populations. Am J Hum Genet. Jan. 2018;102(1):11–26. https://doi.org/10.1016/j.ajhg.2017.11.002.
6. Rogalska ME, Vivori C, Valcárcel J. Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects. Nat Rev Genet. Dec. 2022;1–19. https://doi.org/10.1038/s41576-022-00556-8.
7. Cheung R, et al. A multiplexed assay for exon Recognition reveals that an unappreciated fraction of Rare genetic variants cause large-effect splicing disruptions. Mol Cell. Jan. 2019;73(1):183–94. https://doi.org/10.1016/j.molcel.2018.10.037.e8.
8. Wai HA et al. Jun., 'Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance', *Genet. Med.*, vol. 22, no. 6, Art. no. 6, 2020, https://doi.org/10.1038/s41436-020-0766-9
9. Truty R, et al. Spectrum of splicing variants in disease genes and the ability of RNA analysis to reduce uncertainty in clinical interpretation. Am J Hum Genet. Apr. 2021;108(4):696–708. https://doi.org/10.1016/j.ajhg.2021.03.006.
10. Bournazos AM, et al. Standardized practices for RNA diagnostics using clinically accessible specimens reclassifies 75% of putative splicing variants. Genet Med. Jan. 2022;24(1):130–45. https://doi.org/10.1016/j.gim.2021.09.001.
11. Goyenvalle A et al. Dec., 'Rescue of Dystrophic Muscle Through U7 snRNA-Mediated Exon Skipping', *Science*, vol. 306, no. 5702, pp. 1796–1799, 2004, https://doi.org/10.1126/science.1104297
12. Meulemans L, et al. Skipping nonsense to maintain function: the paradigm of BRCA2 exon 12. Cancer Res. Jan. 2020. https://doi.org/10.1158/0008-5472.CAN-19-2491.

13. Lopez-Perolio I et al. Mar., 'Alternative splicing and ACMG-AMP-2015-based classification of PALB2 genetic variants: an ENIGMA report', *J. Med. Genet.*, p. jmedgenet-2018-105834, 2019, https://doi.org/10.1136/jmedgenet-2018-105834
14. Gonorazky HD, et al. Expanding the boundaries of RNA sequencing as a Diagnostic Tool for Rare mendelian disease. Am J Hum Genet. Mar. 2019;104(3):466–83. https://doi.org/10.1016/j.ajhg.2019.01.012.
15. Mercer TR et al. Nov., 'Targeted RNA sequencing reveals the deep complexity of the human transcriptome', *Nat. Biotechnol.*, vol. 30, no. 1, pp. 99–104, 2011, https://doi.org/10.1038/nbt.2024
16. Davy G et al. Oct., 'Detecting splicing patterns in genes involved in hereditary breast and ovarian cancer', *Eur. J. Hum. Genet. EJHG*, vol. 25, no. 10, pp. 1147–1154, 2017, https://doi.org/10.1038/ejhg.2017.116
17. Adamson SI, Zhan L, Graveley BR. Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. Genome Biol. Jun. 2018;19(1):71. https://doi.org/10.1186/s13059-018-1437-x.
18. Steijger T et al. Dec., 'Assessment of transcript reconstruction methods for RNA-seq', *Nat. Methods*, vol. 10, no. 12, Art. no. 12, 2013, https://doi.org/10.1038/nmeth.2714
19. Workman RE et al. Dec., 'Nanopore native RNA sequencing of a human poly(A) transcriptome', *Nat. Methods*, vol. 16, no. 12, Art. no. 12, 2019, https://doi.org/10.1038/s41592-019-0617-2
20. Glinos DA et al. Aug., 'Transcriptome variation in human tissues revealed by long-read sequencing', *Nature*, vol. 608, no. 7922, Art. no. 7922, 2022, https://doi.org/10.1038/s41586-022-05035-y
21. Treutlein B, Gokce O, Quake SR, Südhof TC. 'Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing', *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 13, pp. E1291-1299, Apr. 2014, https://doi.org/10.1073/pnas.1403244111
22. de Jong LC, et al. Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events. Breast Cancer Res. Nov. 2017;19(1):127. https://doi.org/10.1186/s13058-017-0919-1.
23. Deveson IW, et al. Universal Alternative Splicing of Noncoding Exons. Cell Syst. Feb. 2018;6(2):245–55. https://doi.org/10.1016/j.cels.2017.12.005. .e5.
24. Hardwick SA et al. 'Targeted, High-Resolution RNA Sequencing of Non-coding Genomic Regions Associated With Neuropsychiatric Functions', *Front. Genet.*, vol. 10, 2019, Accessed: Jun. 28, 2023. [Online]. Available: https://www.frontiersin.org/articles/https://doi.org/10.3389/fgene.2019.00309
25. Lagarde J et al. Dec., 'High-throughput annotation of full-length long non-coding RNAs with capture long-read sequencing', *Nat. Genet.*, vol. 49, no. 12, Art. no. 12, 2017, https://doi.org/10.1038/ng.3988
26. Singh M, et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat Commun. Jul. 2019;10(1):3120. https://doi.org/10.1038/s41467-019-11049-4.
27. Hardwick SA, et al. Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. Nat Biotechnol. Jul. 2022;40(7):1082–92. https://doi.org/10.1038/s41587-022-01231-3.
28. O'Leary NA, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. Jan. 2016;44:D733–45. https://doi.org/10.1093/nar/gkv1189. no. D1.
29. Leman R et al. Mar., 'SpliceLauncher: a tool for detection, annotation and relative quantification of alternative junctions from RNAseq data', *Bioinformatics*, vol. 36, no. 5, pp. 1634–1636, 2020, https://doi.org/10.1093/bioinformatics/btz784
30. Love MI, Huber W, Anders S. Moderated estimation of Fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.
31. Dong X et al. Nov., 'Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures', *Nat. Methods*, vol. 20, no. 11, pp. 1810–1821, 2023, https://doi.org/10.1038/s41592-023-02026-3
32. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinforma Oxf Engl. Sep. 2018;34(18):3094–100. https://doi.org/10.1093/bioinformatics/bty191.
33. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. Dec. 2019;20(1):278. https://doi.org/10.1186/s13059-019-1910-1.
34. Quinlan AR, Hall IM. 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinforma. Oxf. Engl.*, vol. 26, no. 6, pp. 841–842, Mar. 2010, https://doi.org/10.1093/bioinformatics/btq033
35. Robinson JT et al. Jan., 'Integrative genomics viewer', *Nat. Biotechnol.*, vol. 29, no. 1, Art. no. 1, 2011, https://doi.org/10.1038/nbt.1754
36. Colombo M et al. Jul., 'Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium', *Hum. Mol. Genet.*, vol. 23, no. 14, pp. 3666–3680, 2014, https://doi.org/10.1093/hmg/ddu075
37. Fackenthal JD et al. Aug., 'Naturally occurring BRCA2 alternative mRNA splicing events in clinically relevant samples', *J. Med. Genet.*, vol. 53, no. 8, pp. 548–558, 2016, https://doi.org/10.1136/jmedgenet-2015-103570
38. Leman R et al. Dec., 'SPiP: Splicing Prediction Pipeline, a machine learning tool for massive detection of exonic and intronic variant effects on mRNA splicing', *Hum. Mutat.*, vol. 43, no. 12, pp. 2308–2323, 2022, https://doi.org/10.1002/humu.24491
39. Castéra L et al. Nov., 'Next-generation sequencing for the diagnosis of hereditary breast and ovarian cancer using genomic capture targeting multiple candidate genes', *Eur. J. Hum. Genet. EJHG*, vol. 22, no. 11, pp. 1305–1313, 2014, https://doi.org/10.1038/ejhg.2014.16
40. Leman R, et al. 2022-RA-935-ESGO Development of an academic genomic instability score for ovarian cancers. Int J Gynecol Cancer. Oct. 2022;32. https://doi.org/10.1136/ijgc-2022-ESGO.596. no. Suppl 2.
41. Richards S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med off J Am Coll Med Genet. May 2015;17(5):405–24. https://doi.org/10.1038/gim.2015.30.
42. Walsh T et al. Dec., 'CRISPR-Cas9/long-read sequencing approach to identify cryptic mutations in BRCA1 and other tumour suppressor genes', *J. Med. Genet.*, vol. 58, no. 12, pp. 850–852, 2021, https://doi.org/10.1136/jmedgenet-2020-107320

Publisher's note